ELSEVIER

Computer Networks 34 (2000) 297–315

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Analysis of shared buffer switches under non-uniform traffic pattern and global flow control

Mahmoud Saleh [a], Mohammed Atiquzzaman [b,*]

[a] *Imam Hussein University, Tehran, Iran*

[b] *Department of Electrical and Computer Engineering, University of Dayton, 300 College Park, Dayton, OH 45469-0226, USA*

Abstract

Shared buffer switches do not suffer from head of line blocking which is a common problem in simple input buffering. Shared buffer switches have previously been studied under uniform and unbalanced traffic patterns. However, due to the complexity of the model, it was not possible to fully explore the performance of such a switch, in the presence of a single hot spot. In this paper, we develop a new model for an ATM switch constructed from shared buffered switching elements, and operating under a hot spot traffic pattern. Hot spot traffic is one of more realistic traffic in ATM switching. The model is validated by comparison with simulation results. The model is used to study the switch performance in terms of throughput, cell delay, cell loss probability and the optimal buffer utilization. Numerical results show that, in the presence of hot spot traffic, shared buffer switches degrade more significantly than switches with dedicated input and/or output buffers. The model can be used by switch designers to optimize the design and performance of switches.

Keywords: Shared buffer switches; Markov chains; Modeling techniques; Performance analysis

1. Introduction

Multistage switches offer a very cost effective architecture for building high speed switches for use in Asynchronous Transfer Mode (ATM) switches for use in Broadband Integrated Services Digital Network. Possible architectures and their performance have been reported in [1–6]. Shared buffer multistage switches use buffer spaces in the switching elements (SE) very efficiently under a uniform traffic pattern [7]. However, an unbalanced traffic in the switch can disrupt the efficiency of a shared buffer SE. For example, if all or a portion of the traffic is directed to a part of the switch instead of being distributed evenly across the switch, the particular part of the switch suffers heavy congestion. The congestion may adversely reduce the performance of the switch and the effectiveness of the shared buffer scheme.

There are several types of non-uniform traffic patterns that can arise in multistage switches. Among them *hot spot* and *point-to-point* (also known as single-source, single-destination) have attracted a lot of attention. The hot spot traffic is a non-uniform traffic pattern consisting of a single output of high access rate (hot spot) superimposed on a background uniform traffic [8]. The hot spot traffic results in a higher throughput at the hot output itself; however, depending on the buffer

* Corresponding author. Tel.: +1-937-229-3183; fax: +1-937-229-4529.

E-mail address: atiq@ieee.org (M. Atiquzzaman).

1389-1286/00/$ - see front matter
PII: S1389-1286(00)00113-4

location and cell forward policies, the overall performance of the switch may face degradation.

Hot spot traffic is a traffic pattern where many sources try to communicate with one destination (hot spot) at the same time [8]. The hot spot traffic pattern could occur in many application areas. For example, many callers may compete to reach a particular subscriber in a telephone network. Other examples of hot spot contention in computer communications are given in [9]. It is, therefore, important to analyze the performance of switches subjected to a hot spot traffic pattern.

Input buffered switches have simple buffer management but suffers from head of line blocking which reduces the throughput [10]. Wu [11] presented an analysis of single input buffered switch under a non-uniform traffic. Kim and Leon-Garcia [12] presented an alternative method for evaluating the performance of input buffered switches. Although their model can be applied to any output distribution, they emphasize single-source, single-destination (SSSD) and superimposed SSSD traffic patterns. They have later extended their model to account for multiple buffers.

A number of authors have studied the performance of output buffered switches [13–16]. Lin and Kleinrock [17] proposed a model to evaluate the performance of multistage switches with output buffering under a hot spot traffic, as well as an extended model for a general traffic distribution. The model uses a decomposition and iterative method to numerically solve the equations. They examine their model with examples for uniform and even-first, odd-second (EFOS) [18] traffic patterns. Although Lin's model may be used for any buffer size, it confines the SE size to 2, and is not suitable for an arbitrary SE size.

The analysis of shared buffer switches under non-uniform traffic patterns was reported by Gianatti and Pattavina [19]. In their model, the outputs of the switch are divided such that a group of outputs are hot and the rest are cold. The number of SEs in the hot group is given by $\log_d N$, where N is the switch size, and d is the size of an SE. For example, for $N = 64$, and $d = 2$, they consider 32 hot and 32 cold outputs. Hence, the model is not suitable for studying switches with a single hot output, where one of the switch outputs becomes more popular than the others. Saleh and Atiquzzaman [20–22] have previously studied the impact of single hot spot traffic on the performance of shared buffer switches using simulation techniques.

Most of the models mentioned above use *local flow* control to control cell movement between stages. In local flow control, a cell can be forwarded to the next stage depending on the buffer occupancy of the next stage at the beginning of an NCC. On the other hand, simultaneous operation of forwarding and receiving cells in a buffer during a switch clock cycle is allowed in *global flow control*. Therefore, global flow control results in a higher throughput and better buffer utilization than local flow control. The *aim of this paper is to study the performance of a multistage switch using local or global flow control, and operating under a hot spot traffic pattern*.

This paper is organized as follows. In Section 2, we describe the modeling assumptions and the single hot spot model. In Section 4, the construction and considerations for a shared buffer switch which is analogous to the one considered for the analysis are discussed. In Section 5, we examine our model with some numerical examples, and compare the results with simulation. Concluding remarks and further possible work are given in Section 6.

2. Shared buffer Delta network

In this section, we will describe the Delta networks, justify our assumptions and notations to be used in the modeling of the network in Section 3. We consider a Delta_d interconnection based switch with N inputs and N outputs and consisting of k stages of $d \times d$ SEs such that $N = d^k$. Starting from the stage where cells enter the network, we number the stages from 1 to k. In a Delta_d interconnection, there exists only one path between each input and output of the network, and each stage of the switch consists of N/d SEs.

In single hot spot traffic, there are i types of SEs at stage i. For example, in Fig. 1, where output 1 is considered as the hot output, the SE types may be

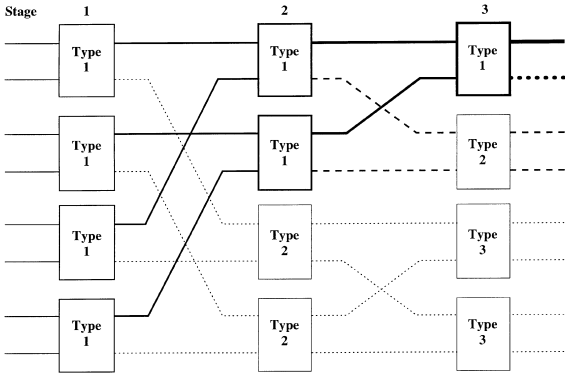

Fig. 1. An 8×8 Delta$_2$ MIN with single hot spot.

labeled as indicated in the figure. Type 1 SE is the only SE in each stage that contains a mixture of hot and cold traffics. SE types greater than 1 contain only cold traffic. However, we still distinguish between different cold type SEs in every stage, because the overall throughput of the SEs is different for the different SE types in the same stage. In Fig. 1, links that carry different traffic mixes are illustrated with different line styles. Hot links and hot SEs carrying hot traffic are shown by thicker solid lines.

2.1. Assumptions

In a shared buffer SE, the buffers must be fast enough to enqueue and dequeue cells during the same network clock cycle (NCC). For the purpose of analysis, we split the process of forwarding and accepting cells in shared buffers during an NCC into two phases [23]. In the *forward* phase, depending on the state of the SE and its downstream SEs, a number of cells may leave the SE, and the switch goes to an *intermediate* state. Following the forward phase is the *receive* phase during which cells offered from the upstream SEs are placed in the buffers, acknowledgments are sent to the upstream SEs, and the SE goes to the *final* state. If the number of arriving cells is greater than the number of available buffers in the SE, a number of cells, equal to the number of available spaces, are selected randomly.

The following assumptions are considered regarding the shared buffer switch and its operation:

- Each SE is of size $d \times d$ and contains B buffers which are accessible by all d inlets and d outlets of the SE.
- The switch operates *synchronously*, i.e. cells are submitted to the switch at the beginning of fixed time intervals which are referred to as network clock cycle (NCC).
- *Destination tag* is used to route a packet. A routing conflict inside the switch is resolved *randomly*, i.e. if two or more cells are destined to the same output, one is chosen at random.
- The arrival of cells at each input of the switch is a *Bernoulli* process, i.e. the probability that a cell arrives during a cycle (ρ) is constant, and successive arrivals are independent of each other.
- The hot traffic is defined as the portion of the input traffic that is *exclusively* destined to the *hot output* and is identified as f_h. Thus, a *hot* SE carries a mixture of hot and cold traffics depending on its location. In the single hot spot case, the hot traffic in an SE passes through only the *hot outlet* of the SE. Therefore, all *cold* SEs contain only *cold* traffic which is uniformly distributed through all outlets in those SEs.
- The probability of a cell arriving at a switch input and being destined to the *hot* output (p_h) or to any single one of the $N-1$ *cold* outputs (p_c) is given by

$$p_h = \rho\left(f_h + \frac{1-f_h}{N}\right), \qquad p_c = \rho\left(\frac{1-f_h}{N}\right), \quad p_h + (N-1)p_c = \rho. \tag{1}$$

- The *state* of an SE whose buffers contain $s = h + c$ cells is represented by a pair (h, c) where h is the number of cells destined to the hot outlet of the SE and c is the number of cells destined to the other $d-1$ cold outlets of the SE.
- Flow control in the switch is implemented by a *backpressure* mechanism which ensures that no cell is lost inside the network. In the case of *local* flow control, a cell leaves an SE if there is a space for it in the corresponding SE at the next stage, at the beginning of an NCC. In *global* flow control, a cell leaves an SE if either there is a space for it in the next stage SE at the beginning of an NCC, or a space becomes available

during the forward phase of the same NCC. An SE acknowledges the receipt of a cell to its upstream SE. Unacknowledged cells contend with other cells in the subsequent cycles.
- There is no *blocking* at the outputs of the network, i.e. an output can always accept a packet.

The possible state transitions in an SE for $d = 2$ and $B = 2$ are illustrated in Fig. 2 for the global flow control policy.

2.2. Notations

The following notation will be used in the models:
- $SE_{i,r}$: an SE of type r at stage i,
- $\pi_{i,r,t}(h1, c1)$: probability that $SE_{i,r}$ is in state $(h1, c1)$ at the beginning of cycle t,
- $\tau_{i,r,t}(h1, c1, h3, c3)$: probability that $SE_{i,r}$ is in state $(h3, c3)$ at the beginning of the receive phase, given that it was in state $(h1, c1)$, at the beginning of the forward phase of cycle t, where $0 \leqslant h1 - h3 \leqslant 1$, and $0 \leqslant c1 - c3 \leqslant d - 1$,
- $\sigma_{i,r,t}(h3, c3, h2, c2)$: probability that $SE_{i,r}$ is in state $(h2, c2)$ at the end of the receive phase of cycle t, given that it was in state $(h3, c3)$ at the beginning of the receive phase of the same cycle, where $h3 \leqslant h2$ and $c3 \leqslant c2$,

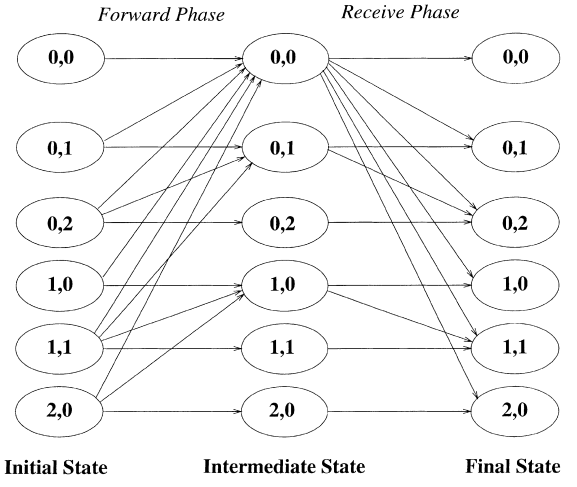

Fig. 2. State diagram of a two-phase switch operation in an SE with $d = 2$, and $B = 2$. Every state is denoted by a pair (h, c) where h is the number of cells destined to the hot outlet of the SE and c is the number of cells destined to the other $d - 1$ cold outlets of the SE.

- $\tilde{\pi}_{i,r,t}(h3, c3)$: probability that $SE_{i,r}$ is in state $(h3, c3)$ at the beginning of the receive phase of cycle t,
- $a_{i,r,t}$: probability that a cell is ready to enter $SE_{i,r}$ at cycle t,
- $b_{i,r,t,x}$: probability that, at cycle t, a successor of $SE_{i,r}$ provides an acknowledgment to type x outlet of the SE, given that a cell was submitted to the successor through outlet x during the same cycle. x is of either a *hot* or *cold* outlet,
- $Y_d(r, c)$: probability that c cells in an SE are destined to r distinct outlets of the SE from a total of d outlets under consideration,
- $u_{i,r,j}$: probability that a cell in $SE_{i,r}$ is destined to its jth outlet, where $1 \leqslant j \leqslant d$.

3. Analysis and modeling of Delta network

In this section, we use the notations and assumptions used in the previous section to develop models for the Delta network under *global* and *local* flow control in Sections 3.1 and 3.2.

3.1. Analysis of an SE with global flow control

We model each SE by a Markov chain representing the distribution of the hot and cold cells stored in the B buffers of the SE. An SE is of type i if it is fed by a type $i - 1$ SE in the previous stage (Fig. 1). It has been shown in [24] that stage i will have i different types of SEs and $i + 1$ different traffic rates at its outlets.

Our modeling approach is based on the iterative solution [25] of a Markov chain system which characterizes the behavior of $SE_{i,r}$ for different i and r in a Delta network. In this approach, if there exists a solution to the system, starting from the resting condition of the switch parameters, the iterative results converge to the steady-state condition of the system.

We represent the state vector of the Markov chain by a row vector $\boldsymbol{\Pi}_{i,r}$ for every type r SE at all of the stages where

$$\boldsymbol{\Pi}_{i,r,t} = [c\pi_{i,r,t}(h, c)], \quad \begin{matrix} h = 0, \ldots, B, \\ c = 0, \ldots, B - h. \end{matrix} \tag{2}$$

Each element of the row vector $\boldsymbol{\Pi}_{i,r}$ is an ordered pair (h,c) in which h is the number of cells in the shared buffers of $SE_{i,r}$ that are destined to the hot outlet of the SE, and c is the number of cells that are destined to the other $d-1$ outlets of the SE. It is, thus, obvious that the total number of cells in the shared buffers for $\pi_{i,r}(h,c)$ is $h+c$.

Based on the imaginary intermediate state assumption, the probability that the state of $SE_{i,r}$ goes to an intermediate state $(h3,c3)$ is equal to the SE being in state $\pi_{i,r,t}(h1,c1)$ and a τ transition from $(h1,c1)$ to $(h3,c3)$ taking place for all possible $(h1,c1)$ states. In other words,

$$\tilde{\pi}_{i,r,t}(h3,c3) = \sum_{h1=0}^{B} \sum_{c1=0}^{B-h1} \pi_{i,r,t}(h1,c1)\tau_{i,r,t} \times (h1,c1,h3,c3). \quad (3)$$

The limits of the summations ensure that $h1+c1 \leqslant B$. Similarly, the probability of the transition to the initial state $(h2,c2)$ at time $t+1$ (which is equivalent to the final state at time t) is equal to the SE being in the intermediate state $(h3,c3)$, and a σ transition from $(h3,c3)$ to $(h2,c2)$ taking place for all possible $(h3,c3)$ states. In other words,

$$\pi_{i,r,t+1}(h2,c2) = \sum_{h3=0}^{B} \sum_{c3=0}^{B-h3} \tilde{\pi}_{i,r,t}(h3,c3)\sigma_{i,r,t} \times (h3,c3,h2,c2). \quad (4)$$

In this case, too, the limits of the summations are such that $h3+c3 \leqslant B$.

To avoid unnecessary complexity in our notation and since we are interested in the steady-state condition of the network, we drop the t subscripts in all NCC dependent formulas in the rest of this paper. This does not affect the interpretation of any formulas, since, for instance, the value of a formula at time t and $t+1$ is the same, for sufficiently large t. However, in calculation, one should be aware that such formulas are actually time dependent.

Transition $\tau_{i,r}(h1,c1,h3,c3)$ in Eq. (3) takes place if $h1-h3$ cells leave $SE_{i,r}$ from its hot outlet, and $c1-c3$ cells leave the SE from its $d-1$ cold outlets. Since all hot traffic passes through only the hot outlet, at most one cell can leave from the hot outlet, depending on whether it is accepted in the next stage. Thus, the probability that a cell leaves $SE_{i,r}$ through its hot outlet has a binomial distribution form

$$\beta(h1-h3, \min(1,h1), b_{i,r,\text{hot}}), \quad (5)$$

where β is the shorthand notation for a binomial distribution

$$\beta(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k}. \quad (6)$$

Note that at most one cell can leave through any outlet during each NCC. In Eq. (5) the probability for the cases where $h1-h3>1$ is 0. This is implicit in the binomial distribution where $\binom{n}{k}=0$ for $k>n$.

On the other hand, the probability that $c1-c3$ leave the $SE_{i,r}$ through other $d-1$ cold outlets depends on:

1. the $c1$ cells are destined to exactly how many outlets, and
2. the probability that $c1-c3$ cells pass through those outlets.

Note that at most one cell leaves an SE in any NCC. Thus, the probability that $c1-c3$ leave the $SE_{i,r}$ through other $d-1$ cold outlets implies that $c1$ cells are destined to at least $c1-c3$ different outlets. Considering this case for all possible number of outlets, we have

$$\sum_{l=c1-c3}^{d-1} Y_{d-1}(l,c1)\beta(c1-c3,l,b_{i,r,\text{cold}}). \quad (7)$$

The lower limit of $l=c1-c3$ is due to the fact that $c1$ packets should be destined to at least $c1-c3$ distinct outlets so that the same number of cells as $c1-c3$ are able to leave the SE. The upper limit reflects the fact that there are at most $d-1$ cold outlets in each SE. Since, the probabilities in Eqs. (5) and (7) are independent, $\tau_{i,r}(h1,c1,h3,c3)$ is equal to the product of the two equations:

$$\tau_{i,r}(h1,c1,h3,c3) = \beta(h1-h3, \min(1,h1), b_{i,r,\text{hot}}) \times \sum_{l=c1-c3}^{d-1} Y_{d-1}(l,c1)\beta(c1-c3,l,b_{i,r,\text{cold}}), \quad (8)$$

where d is the number of inlets and outlets of an SE.

Since we assume that the cold traffic of an SE is distributed uniformly over all $d-1$ outlets of an SE, $Y_d(c,s)$ may be obtained by Bianchi and Turner [26]:

$$Y_d(c,s) = \binom{d}{c} \frac{\gamma(s-c,c)}{\gamma(s,d)}, \tag{9}$$

where

$$\gamma(s,d) = \binom{s+d-1}{s}. \tag{10}$$

Y_d is independent of SE type and stage number. It can, therefore, be calculated once and used for subsequent calculations in order to reduce computing time.

The probability that a cell which is sent through an outlet of type x of $SE_{i,r}$ is accepted by its next stage, $b_{i,r,x}$, depends on the stage and type of the SE:

1. $i = k$. Since there is no blocking at the outputs of the network, the probability $b_{i,r,x}$ of acceptance of an offered cell at stage k (the last stage) is equal to 1.
2. $i < k$. For all stages except the last stage we shall consider three different cases:

(a) An offered cell to a particular outlet of $SE_{i,r}$ is definitely accepted by its successor SE, if there are at least d buffers in the successor SE. The probability of this case will be

$$\sum_{h3=0}^{B-d} \sum_{c3=0}^{B-d-h3} \tilde{\pi}_{i+1,s}(h3,c3). \tag{11}$$

The upper limits and lower limits of the summations ensure that $h3 + c3 \leqslant B - d$, so that there are at least d empty buffers in the shared buffer SE. Subscript s used in Eqs. (11)–(18) denotes the type of SE which should be considered at the next stage. The type in the next stage is determined as

$$s = \begin{cases} 1, & i = 1 \text{ and } x = 1, \\ r+1, & r > 1 \text{ or } (r = 1 \text{ and } x \neq 1). \end{cases} \tag{12}$$

(b) An offered cell to a particular outlet of $SE_{i,r}$ is also accepted if the total number of cells that are offered to other $d-1$ inlets of the successor SE is *strictly* less than the number of available buffers in that SE. If w_h is the number of packets destined to the hot outlet at stage $i+1$ and w_c is the number of cells destined to the cold outlets at stage $i+1$, both through other $d-1$ outlets of $SE_{i,r}$, the probability that w_h cells are destined to the hot outlet and w_c cells are destined to $d-1$ cold outlets forms a multinomial distribution. Thus, b_i, in this case, will be equal to

$$\sum_{h3=0}^{B} \sum_{c3=0}^{B} \tilde{\pi}_{i+1,s}(h3,c3) \times \sum_{w_h=0}^{d-1} \sum_{w_c=0}^{d-1} \mu(a_{i+1,s}u_{i+1,s,\text{hot}}, d-1, w_h, w_c). \tag{13}$$

$$B-d+1 \leqslant h3+c3 \leqslant B-1 \qquad w_h+w_c \leqslant B-(h3+c3)-1$$

For example, if $d = 2$ and $B = 2$, Eq. (13) will reduce to

$$\left[\tilde{\pi}_{i+1,s}(1,0) + \tilde{\pi}_{i+1,s}(0,1)\right](1 - a_{i+1,s}), \tag{14}$$

which means that a cell which has been offered to $SE_{i+1,s}$ will be accepted if the SE is in one of two intermediate states $(1,0)$ or $(0,1)$, and no cell is offered at the other inlet of $SE_{i+1,s}$.

$\mu(a,u,d,h,c)$ is the multinomial distribution of h and c from a total of d

$$\mu(a,u,d,h,c) = \frac{d!}{h!c!(d-h-c)!}(a.u)^h \times [a(1-u)]^c (1-a)^{d-h-c}. \tag{15}$$

(c) If $w_h + w_c$ (having the same definition as in the previous case) is greater than or equal to the number of available buffers, the probability that an acknowledgment is received at the output under consideration is a fraction of the previous case depending on the value of w_h and w_c.

$$\sum_{h3=0}^{B} \sum_{c3=0}^{B} \tilde{\pi}_{i+1,s}(h3,c3) \times \sum_{w_h=0}^{d-1} \sum_{w_c=0}^{d-1} \mu(a_{i+1,s}u_{i+1,s,\text{hot}}, d-1, w_h, w_c) \times \frac{B-(h3+c3)}{w_h+w_c+1}. \tag{16}$$

$$B-d+1 \leqslant h3+c3 \leqslant B-1 \qquad B-(h3+c3) \leqslant w_h+w_c \leqslant d-1$$

For example, if $d = 2$ and $B = 2$, Eq. (16) reduces to

$$(\tilde{\pi}_{i+1,s}(1,0)+\tilde{\pi}_{i+1,s}(0,1)) \times\left(\frac{1}{2}a_{i+1,s}u_{i,s,\mathrm{hot}}+\frac{1}{2}a_{i+1,s}(1-u_{i,s,\mathrm{hot}})\right), \quad (17)$$

which means that – in this case of $b_{i,r,x}$ – an offered cell to $\mathrm{SE}_{i+1,s}$ will be acknowledged with the probability $\frac{1}{2}$ if the SE is in either the intermediate states $(1,0)$ or $(0,1)$, and another inlet of $\mathrm{SE}_{i+1,s}$ definitely has a cell which is destined either to the hot outlet or the cold outlets of $\mathrm{SE}_{i+1,s}$.

$(B-(h3+c3))/(w_\mathrm{h}+w_\mathrm{c}+1)$ in Eq. (16) assumes that cells at an SE are accepted in the buffers at random, regardless of whether they are destined to the hot outlet or any one of the cold outlets.

Since all cases are independent of each other, $b_{i,r,x}$ is the sum of all three cases

$$\begin{aligned} b_{i,r,x} &= \sum_{h3=0}^{B-d}\sum_{c3=0}^{B-d-h3}\tilde{\pi}_{i+1,s}(h3,c3) \\ &+ \sum_{h3=0}^{B}\sum_{c3=0}^{B}{}^{B-d+1\leqslant h3+c3\leqslant B-1}\tilde{\pi}_{i+1,s}(h3,c3) \\ &\times\left[\sum_{w_\mathrm{h}=0}^{d-1}\sum_{w_\mathrm{c}=0}^{d-1}{}^{w_\mathrm{h}+w_\mathrm{c}\leqslant B-(h3+c3)-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_\mathrm{h},w_\mathrm{c})\right. \\ &+ \sum_{w_\mathrm{h}=0}^{d-1}\sum_{w_\mathrm{c}=0}^{d-1}{}^{B-(h3+c3)\leqslant w_\mathrm{h}+w_\mathrm{c}\leqslant d-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_\mathrm{h},w_\mathrm{c}) \\ &\left.\times\frac{B-(h3+c3)}{w_\mathrm{h}+w_\mathrm{c}+1}\right]. \end{aligned} \quad (18)$$

In Eq. (4), $\sigma_i(h3,c3,h2,c2)$ is the probability that a transition from intermediate state $(h3,c3)$ to final state $(h2,c2)$ takes place. $\sigma_i(h3,c3,h2,c2)$ depending on the final state $(h2,c2)$ consists of two cases.

(1) $h2+c2<B$. This means that after receiving $(h2+c2)-(h3+c3)$ the buffers are not yet completely full. In this case all of $(h2+c2)-(h3+c3)$ cells that were ready to enter $\mathrm{SE}_{i,r}$ have already entered the SE, since there was room for all of them. Among the cells that have entered the SE, there are $h2-h3$ cells that are destined to the hot outlet of the SE and $c2-c3$ cells that are destined to other $d-1$ cold outlets of the SE. Therefore, $\sigma_{i,r}$ in this case has a multinomial distribution of the following form:

$$\begin{aligned} &\sigma_{i,r}(h3,c3,h2,c2) \\ &\quad=\mu(a_{i,r},u_{i,r,\mathrm{hot}},d,h2-h3,c2-c3), \\ &\quad h2+c2<B. \end{aligned} \quad (19)$$

(2) $h2+c2=B$. This means that the buffers are all full after the transition took place. However, it does not mean that there were only $(h2-h3)+(c2-c3)$ cells which were ready to enter the SE. In fact, the number of cells that may have been ready to enter SE_i can be any number between $(h2-h3)+(c2-c3)$ and d. In other words, $\sigma_{i,r}$ for the case of $h2+c2=B$ will be

$$\begin{aligned} &\sigma_{i,r}(h3,c3,h2,c2) \\ &\quad=\sum_{w_\mathrm{h}=h2-h3}^{d-(c2-c3)}\sum_{w_\mathrm{c}=c2-c3}^{d-w_\mathrm{h}}\frac{\binom{w_\mathrm{h}}{h2-h3}\binom{w_\mathrm{c}}{c2-c3}}{\binom{w_\mathrm{h}+w_\mathrm{c}}{c2-c3+h2-h3}} \\ &\qquad\times\mu(a_{i,r},u_{i,r,\mathrm{hot}},d,w_\mathrm{h},w_\mathrm{c}), \\ &\quad h2+c2=B. \end{aligned} \quad (20)$$

For example, if $d=2$ and $B=2$, and we would like to consider different cases for $\sigma_{i,r}(0,1,1,1)$, the following cases during the current NCC are possible:

- Only one cell was offered to $\mathrm{SE}_{i,r}$ and it was destined to the hot outlet.
- Two cells were offered to $\mathrm{SE}_{i,r}$, both destined to the hot outlet, and one of them was accepted.
- Two cells were offered to $\mathrm{SE}_{i,r}$, one destined to the hot outlet and one to the cold outlet, and only the one which was destined to the hot outlet was accepted.

Note that in the hot spot model, when a cell enters an SE it is not important which inlet it came from. The state of the SE only represents which outlet it is destined to.

The coefficient

$$\frac{\binom{w_h}{h2-h3}\binom{w_c}{c2-c3}}{\binom{w_h+w_c}{c2-c3+h2-h3}}$$

in Eq. (20) is the probability that $h2-h2$ cells are selected from w_h cells destined to the hot outlet and $c2-c3$ cells are selected from w_c cells destined to the cold outlets, given that a total of $(c2-c3)+(h2-h3)$ cells are accepted to $SE_{i,r}$, out of w_h+w_c cells that were offered to that SE. The limits of the summations are determined as follows. The number of cells that were offered to $SE_{i,r}$ and are destined to the hot outlet of the SE (w_h) must be greater or equal to $h2-h3$ (the actual number of cells that entered $SE_{i,r}$ and are destined to its hot outlet). On the other hand, w_h is limited to the number of inlets of the SE minus $c2-c3$ (the actual number of cells that enter $SE_{i,r}$ and are destined to its $d-1$ cold outlets). The lower bound of w_c is justified as similar to that of w_h. The upper bound of w_c is limited to $d-w_h$ which is the SE size minus the number of cells that are offered to $SE_{i,r}$.

Putting both cases in one formula, $\sigma_{i,r}(h3,c3,h2,c2)$ is the following:

$$\sigma_{i,r}(h3,c3,h2,c2) = \begin{cases} \mu(a_{i,r},u_{i,r,\mathrm{hot}},d,h2-h3,c2-c3), \\ \qquad h2+c2<B, \\ \displaystyle\sum_{w_h=h2-h3}^{d-(c2-c3)}\sum_{w_c=c2-c3}^{d-w_h}\frac{\binom{w_h}{h2-h3}\binom{w_c}{c2-c3}}{\binom{w_h+w_c}{c2-c3+h2-h3}} \\ \qquad \times\mu(a_{i,r},u_{i,r,\mathrm{hot}},d,w_h,w_c), \\ \qquad h2+c2=B. \end{cases} \tag{21}$$

The probability that a cell is ready to enter $SE_{i,r}$, $a_{i,r}$, depends on i and r. There are three different cases which should be distinguished:

1. $i=1$. For $i=1$, the $a_{i,r}$ is equal to the probability that a cell is offered to the Delta interconnection at a particular NCC. We have assumed that all inputs are independent and the probability that there is a packet available at any input is equal to ρ

$$a_{i,r}=\rho, \quad i=1. \tag{22}$$

2. $i>1$ *and* $r=1$. This case represents the hot SEs at all stages except the first. All cells that are entering this type of SE are coming from the hot outlet of type 1 SEs (switching elements carrying a mixture of hot and cold traffics). Thus, the probability that a cell enters $SE_{i,r}$ depends on whether there is at least one cell at any one of $SE_{i-1,r}$s. Since $a_{i,r}$ is the same for all inlets, we can write

$$a_{i,r}=1-\sum_{j=0}^{B}\pi_{i-1,r}(0,j), \quad i>1,\ r=1. \tag{23}$$

The summation in Eq. (23) is the probability that no cell is destined to the hot outlet of $SE_{i-1,r}$. Therefore, the probability that at least one cell is destined to the hot outlet of that SE is equal to the subtraction of the sum from one.

3. $i>1$ *and* $r>1$. This case represents all other SEs, in stages other than the first, which carry only cold traffic. In this case the probability that there is at least one cell which is ready to enter $SE_{i,r}$ depends on whether there is at least one cell in any one of $SE_{i-1,r-1}$s which is destined to one of the cold outlets of that SE. Note that $a_{i,r}$ is the same for every inlet at stage i which falls in this category. Thus, we can write

$$a_{i,r}=\sum_{l=1}^{B}\sum_{j=0}^{B-l}\pi_{i-1,r-1}(j,l)\left(\frac{1-\gamma(s,d-2)}{\gamma(s,d-1)}\right), \quad i>1,\ r>1, \tag{24}$$

where $\gamma(s,d)$ is as defined in Eq. (10). Therefore, the combined equation for $a_{i,r}$ will be the following:

$$a_{i,r}=\begin{cases} \rho, & i=1, \\ 1-\sum_{j=0}^{B}\pi_{i-1,r}(0,j), & i>1, \\ & r=1, \\ \sum_{l=1}^{B}\sum_{j=0}^{B-l}\pi_{i-1,r-1}(j,l)\left(1-\frac{\gamma(s,d-2)}{\gamma(s,d-1)}\right), & i>1, \\ & r>1. \end{cases} \tag{25}$$

The probability that a cell in $SE_{i,r}$ is destined to outlet j of the SE, $u_{i,r,j}$, is determined by

$$u_{i,r,j}=\frac{e_{i,r,j}}{m_{i,r}}, \tag{26}$$

where $m_{i,r}$ is the sum of all $e_{i,r,j}$ for all j in $\mathrm{SE}_{i,r}$

$$m_{i,r} = \sum_{l=1}^{d} e_{i,r,l}. \tag{27}$$

For the last stage, the probability $e_{k,r,j}$ that a cell is referencing a hot or cold output is simply calculated by

$$e_{k,r,j} = \begin{cases} p_{\mathrm{h}}, & r = 1 \text{ and } j = 1, \\ p_{\mathrm{c}} & (r = 1 \text{ and } j > 1) \text{ or } r > 1, \end{cases} \tag{28}$$

where $k = \log_d N$. For $i < k$, $e_{i,r,j}$ is calculated as follows:

$$e_{i,r,j} = \begin{cases} m_{i+1,r}, & j = 1, \\ m_{i+1,r+1}, & 1 < j \leqslant d. \end{cases} \tag{29}$$

In [27], we have derived $e_{i,r,j}$ and $m_{i,r}$ in a different way which avoids recursive calculation. Nevertheless, both techniques are correct and give the same results. However, the derivation here is more concise and easier to understand.

Since $u_{i,r,j}$ values are independent of time, it is more efficient to calculate $e_{i,r,j}$, $m_{i,r}$, and $u_{i,r,j}$ values once, store them in an appropriate data structure and use them for the rest of the calculations. Starting the calculations from the last stage is advantageous, because some already calculated values for next stages can be directly used in the calculations for the current stage. Thus, the calculation takes place with no recursive calls.

An example of calculation of $u_{i,r}$ is given in Fig. 3 for $N = 4$, $\rho = 0.8$, and $f_{\mathrm{h}} = 0.2$.

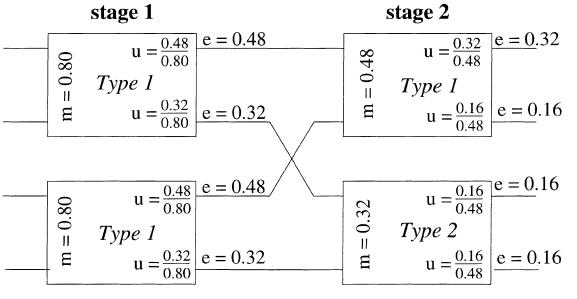

Fig. 3. An example of $u_{i,r,j}$, $m_{i,r}$, and $e_{i,r,j}$ values for a Delta interconnection with $N = 4, d = 2, \rho = 0.8$, and $f_{\mathrm{h}} = 0.2$. Subscripts are omitted for readability.

3.2. Analysis of an SE with local flow control

The model from a multistage switch using global flow control was described in the previous section. In this section, we show how to model a Delta interconnection using local flow control.

The state probabilities of $\mathrm{SE}_{i,r}$ in local flow control may be described by a Markov chain with the following description:

$$\pi_{i,r}(h2, c2) = \sum_{h1=0}^{B} \sum_{c1=0}^{B-h1} \pi_{i,r}(h1, c1)\theta_{i,r} \times (h1, c1, h2, c2), \tag{30}$$

where $\theta_{i,r}(h1, c1, h2, c2)$ is the probability that an SE is in state $(h2, c2)$, in the current cycle, given that it was in state $(h1, c1)$ in the previous cycle

$$\theta_{i,r}(h1, c1, h2, c2) = \sum_{w=\max(0,h2-h1)}^{w_m} \sum_{s=\max(0,c2-c1)}^{s_m} \sigma_{i,r}(h1, c1, h1 + w, c1 + s) \times \tau_{i,r}(h1, c1, h2 - w, c2 - s), \tag{31}$$

where the upper limits w_m and s_m of the summation are as follows:

$$w_m = \min(d, h2, B - (h1 + c1), h2 - h1 + 1),$$
$$s_m = \min(d, c2, B - (h1 + c1) - w, d - 1 + c2 - c1).$$

In local flow control, the process of forwarding cells from an SE is the same as in global flow control. However, the process of accepting cells at the inlets of the SE is different from global flow control. Therefore, we need to derive different formulas for the variables which are affected. Fortunately, we only need to change $b_{i,r}$ and $\sigma_{i,r}$ equations. The rest of the equations remain the same as in global flow control.

As in global flow control, the probability $b_{i,r,x}$ of acceptance of an offered cell at the last stage is equal to 1.

For all stages except the last stage we shall consider three different cases:

1. An offered cell to a particular outlet of $\mathrm{SE}_{i,r}$ is definitely accepted by its successor SE, if there are at least d buffers in the successor SE. The probability of this case will be

$$\sum_{h1=0}^{B-d}\sum_{c1=0}^{B-d-h1}\pi_{i+1,s}(h1,c1). \tag{32}$$

The upper limits and lower limits of the summations ensure that $h1+c1\leqslant B-d$, so that there are at least d empty buffers in the shared buffer SE.

2. An offered cell to a particular outlet of $SE_{i,r}$ is also accepted if the total number of cells that are offered to other $d-1$ inlets of the successor SE is *strictly* less than the number of available buffers in that SE. If w_h is the number of packets destined to the hot outlet at stage $i+1$ and w_c is the number of cells destined to the cold outlets at stage $i+1$, both through other $d-1$ outlets of $SE_{i,r}$, the probability that w_h cells are destined to the hot outlet and w_c cells are destined to $d-1$ cold outlets forms a multinomial distribution. Thus, b_i, in this case, will be equal to

$$\sum_{h1=0}^{B}\sum_{c1=0}^{B}{}^{B-d+1\leqslant h1+c1\leqslant B-1}\pi_{i+1,s}(h1,c1) \times \sum_{w_h=0}^{d-1}\sum_{w_c=0}^{d-1}{}^{w_h+w_c\leqslant B-(h1+c1)-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_h,w_c). \tag{33}$$

For example, if $d=2$ and $B=2$, Eq. (33) will reduce to

$$[\pi_{i+1,s}(1,0)+\pi_{i+1,s}(0,1)](1-a_{i+1,s}), \tag{34}$$

which means that a cell which has been offered to $SE_{i+1,s}$ will be accepted if the SE is in one of two initial states $(1,0)$ or $(0,1)$, and no cell is offered at the other inlet of $SE_{i+1,s}$. Subscript s used in Eqs. (32)–(38) denotes the type of SE which should be considered at the next. The type in the next stage is determined as

$$s=\begin{cases}1, & i=1 \text{ and } x=1,\\ r+1, & r>1 \text{ or } (r=1 \text{ and } x\neq 1).\end{cases} \tag{35}$$

3. If w_h+w_c (having the same definition as in the previous case) is equal or more than the number of available buffers, the probability that an acknowledgment is received at the output under consideration is a fraction the previous case depending on the value of w_h and w_c

$$\sum_{h1=0}^{B}\sum_{c1=0}^{B}{}^{B-d+1\leqslant h1+c1\leqslant B-1}\pi_{i+1,s}(h1,c1) \times \sum_{w_h=0}^{d-1}\sum_{w_c=0}^{d-1}{}^{B-(h1+c1)\leqslant w_h+w_c\leqslant d-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_h,w_c) \times\frac{B-(h1+c1)}{w_h+w_c+1}. \tag{36}$$

For example, if $d=2$ and $B=2$, Eq. (36) reduces to

$$(\pi_{i+1,s}(1,0)+\pi_{i+1,s}(0,1)) \times\left(\frac{1}{2}a_{i+1,s}u_{i,s,\mathrm{hot}}+\frac{1}{2}a_{i+1,s}(1-u_{i,s,\mathrm{hot}})\right), \tag{37}$$

which means that – in this case of $b_{i,r,x}$ – an offered cell to $SE_{i+1,s}$ will be acknowledged with the probability $\frac{1}{2}$ if the SE is in either the initial states $(1,0)$ or $(0,1)$, and another inlet of $SE_{i+1,s}$ definitely has a cell which is destined either to the hot outlet or the cold outlet of $SE_{i+1,s}$.

$(B-(h1+c1))/(w_h+w_c+1)$ in Eq. (36) assumes that cells at an SE are accepted in the buffers at random, regardless of whether they are destined to the hot outlet or any one of the cold outlets.

Since all cases are independent of each other, $b_{i,r,x}$ for stages other than the last is the sum of all three cases

$$\begin{aligned}b_{i,r,x}=&\sum_{h1=0}^{B-d}\sum_{c1=0}^{B-d-h1}\pi_{i+1,s}(h1,c1)\\ &+\sum_{h1=0}^{B}\sum_{c1=0}^{B}{}^{B-d+1\leqslant h1+c1\leqslant B-1}\pi_{i+1,s}(h1,c1)\\ &\times\left[\sum_{w_h=0}^{d-1}\sum_{w_c=0}^{d-1}{}^{w_h+w_c\leqslant B-(h1+c1)-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_h,w_c)\right.\\ &+\sum_{w_h=0}^{d-1}\sum_{w_c=0}^{d-1}{}^{B-(h1+c1)\leqslant w_h+w_c\leqslant d-1}\mu(a_{i+1,s}u_{i+1,s,\mathrm{hot}},d-1,w_h,w_c)\\ &\left.\times\frac{B-(h1+c1)}{w_h+w_c+1}\right].\end{aligned} \tag{38}$$

Comparison of Eq. (38) with Eq. (18) shows that the only difference between them is that in Eq. (18) $b_{i,r,x}$ depends on the intermediate states of $SE_{i+1,s}$, whereas Eq. (38) depends on the initial states. A similar analogy exists between $\sigma_{i,r}$ in the global flow control equations and $\sigma_{i,r}$ in the case of local flow control. Therefore, we do not elaborate on the $\sigma_{i,r}$ in the local flow control case.

$\sigma_{i,r}$ in local flow control is given by

$$\sigma_{i,r}(h1,c1,h2,c2) = \begin{cases} \mu(a_{i,r},u_{i,r,\mathrm{hot}},d,h2-h1,c2-c1), \\ \qquad h2+c2<B, \\ \displaystyle\sum_{w_\mathrm{h}=h2-h1}^{d-(c2-c1)} \sum_{w_\mathrm{c}=c2-c1}^{d-w_\mathrm{h}} \frac{\binom{w_\mathrm{h}}{h2-h1}\binom{w_\mathrm{c}}{c2-c1}}{\binom{w_\mathrm{h}+w_\mathrm{c}}{c2-c1+h2-h1}} \\ \qquad \times\mu(a_{i,r},u_{i,r,\mathrm{hot}},d,w_\mathrm{h},w_\mathrm{c}), \\ \qquad h2+c2=B. \end{cases} \tag{39}$$

3.3. Order of calculation in the hot spot models

The Markov chain models in Sections 3.1 and 3.2 are numerically calculated using the method first described in [25] for multistage networks. We describe the order of calculation which we used for the numerical solution to the Markov chain equations for global flow control. The solution for local flow control policy is very similar to the case of global flow control.

For hot spot model and global flow control policy, the dependency between different stochastic variables is as follows:

1. *Time independent values*

(a) $Y_d(c,s)$ is the probability that s cells are destined to c different outlets in an SE. The values are independent of time and stage number or switch type. Thus, it is recommended that $Y_d(c,s)$ is calculated based on Eq. (9) for all $0 \leqslant c \leqslant d$ and $0 \leqslant s \leqslant B$, and the values are stored in a table and used whenever necessary. This requires a two-dimensional matrix of size $(d+1)\times(B+1)$.

(b) $u_{i,r,j}$ is the probability that a cell in the $SE_{i,r}$ is destined to outlet j of the SE which is based on $e_{i,r,j}$ which can be directly or recursively [27] calculated. $u_{i,r,j}$ values may be calculated once, stored in appropriate data structures, and used as necessary in the following calculations.

2. *Time dependent variables*

Table 1 summarizes the dependencies of all time dependent variables in the solution of the Markov chain for the hot spot model. Note that only stage number and time subscript t are shown for each variable. Other subscripts such as SE type or SE outlet number are as in the equations described before.

Table 1 helps in finding the correct order of the numerical solution to the Markov chain equations. We used a C program to solve the Markov chain system. The program reads the SE size d, buffer size B, and number of stages k from the input, and dynamically allocates the necessary space needed to internally represent the multistage Delta network. Then it reads in the values of the input rate ρ, and hot spot value f_h. After allocating the right data structure for each variable, calculation of various variables is performed in the following order:

(a) *Initialization of the data structures*. All data structures are assigned with the appropriate initial (rest) conditions. For example, the initial value for $\pi_{i,r}$ vector, that contains the initial state probabilities at $SE_{i,r}$, and for $\tilde{\pi}_{i,r}$ vector, that contains the intermediate values, should indicate that there is no cell in the shared buffer at any SE.

$$\pi_{i,r,t=0} = \tilde{\pi}_{i,r,t=0} = [1\ 0\ \cdots\ 0]. \tag{40}$$

The initial value of $b_{i,r,j}$ is 1 for all outlets of $SE_{i,r}$ and for all $SE_{i,r}$. The initial value for $a_{i,r}$ is equal to ρ for $i=1$, and 0 otherwise. The initial values of $\sigma_{i,r}$ and $\tau_{i,r}$ matrices could be any arbitrary values, since they are replaced with the correctly calculated values later. We initialize

Table 1
Dependency of time dependent variables in hot spot model

Variable name	Depends on
$b_{i,t}$	$\tilde{\pi}_{i+1,t}$, $\sigma_{i+1,t}$, $a_{i+1,t}$, u_{i+1}
$a_{i,t}$	$\pi_{i-1,t}$
$\tau_{i,t}$	Y_d, $b_{i,t}$
$\tilde{\pi}_{i,t}$	$\pi_{i,t}$, $\tau_{i,t}$
$\sigma_{i,t}$	$a_{i,t}$, u_i
$\pi_{i,t}$	$\tilde{\pi}_{i,t-1}$, $\sigma_{i,t-1}$

all such data structures with 0 for debugging purposes.

(b) *Iterative calculation of the variables*. After proper initialization of the variables, the iterative calculation of the Markov chain equations is performed in the order as described in the main loop in Algorithm 1.

Our experience shows that starting from the initial condition of the Markov chain system, the system converges to its steady-state values. However, the rate of convergence depends on the hot spot value, SE size and switch size. The number of iterations vary from $2k$ to a few hundred times. It is generally necessary to confine the number of iterations or the acceptable convergence to an upper limit. In our approach, we let the iteration loop be executed at least $k+1$ times, since this is the minimum cycles that allows the equations at all stages to be assigned to values different from the initial values. After that, the iteration loop is terminated based on Algorithm 2. In this algorithm the state vector $\boldsymbol{\Pi}_{i,r,t}$ is compared with $\boldsymbol{\Pi}_{i,r,t-1}$ in all stages.

Algorithm 2 proves quite satisfactory in examining whether the Markov chain system has converged to its steady-state condition. The execution time for the algorithm is negligible compared to the calculation time of the Markov chain equations.

Algorithm 1. Order of calculation of different variables in the hot-global-ack model

```
    for i = k to 1 do {for all stages in the multistage network}
      calculate Y_d(c, s) for all values of 0 ≤ c ≤ d and 0 ≤ s ≤ B, and store them in the
      relevant data structure.
      for r = 1 to i do {for all SE types in stage i}
        calculate u_{i,r}
 5:   end for
    end for
    while not steady-state condition do {repeat the loop while not reached the steady state}
      for i = k to 1 do {for all stages in the multistage network}

10:     for r = 1 to i do {for all SE types in stage i}

          for j = 1 to d do {for all outlets in SE_{i,r}}
            calculate new b_{i,r,j}.
          end for
15:       calculate new a_{i,r}.
        end for
        for r = 1 to i do {for all SE types in stage i}
          calculate π̃_{i,r} based on the initial or previous values of π_{i,r}.
        end for
20:     for r = 1 to i do {for all SE types in stage i}
          calculate σ_{i,r} for all states in the shared buffer.
          calculate τ_{i,r} for all states in the shared buffer.
        end for
        for r = 1 to i do {for all SE types in stage i}
25:       calculate the new values of π_{i,r} based on values calculated during the current
          cycle.
        end for
      end for
    end while
```

Algorithm 2. Criteria for terminating the iteration loop in the hot spot model

```
terminate ← TRUE {terminate is a logical flag which is initially TRUE.}
i ← 1 {i is the stage counter.}
while (i ≤ k) ∧ terminate = TRUE do
  r ← 1 {r is the SE type counter in each stage.}
  while (r ≤ i) ∧ terminate = TRUE do
    if (Σ_{h=0}^{B} Σ_{c=0}^{B−h} [π_{i,r,t}(h, c) − π_{i,r,t−1}(h, c)]^2) > 10^{−4} then
      terminate = FALSE
    end if
    r ← r + 1
  end while
  i ← i + 1
end while
if terminate = TRUE then
  terminate the iteration loop in Algorithm 1.
end if
```

It should be noted that data structures for $\sigma_{i,r}$ and $\tau_{i,r}$ become huge for large B. For example, for $B=50$, one would normally need 51^4 memory units of type double in C for each $\sigma_{i,r}$ and $\tau_{i,r}$, which corresponds to 51 megabytes of memory on a Unix system! However, most of the space is essentially useless, since some transitions are impossible in an SE. For example, for $B=50$, transition $\tau_{i,r}(50,50,50,50)$ is not possible, since there are no more than 50 spaces available in the shared buffer. $\sigma_{i,r}$ and $\tau_{i,r}$ matrices are, in fact, very sparse for large B, and must be dealt with using advanced data structures such as hash tables. We have used a hash table technique to tackle this problem in solving the hot spot model. For example, the number of allowable transitions of $\sigma_{i,r}$ for $B=4$ and $B=50$ is 1660 which requires 13 280 memory bytes for storage. The calculation overhead of hash table lookup and store for our implementation is about 5% of the total run time.

3.4. Performance evaluation

Analytical models for a multistage shared buffer switch for local and global flow control policies have been developed in the previous section. In this section, we define the parameters which will be used to obtain the performance of such a switch. In steady-state condition of the switch, the throughput, cell loss, and delay of various SE types can be computed.

Throughput. Throughput at an SE outlet is defined as the probability of a cell leaving that outlet during an NCC. In the single hot spot model, the hot and cold outlets of an SE have different

throughputs, whereas the throughputs of all cold outlets of an SE are the same.

The throughput of the hot outlet of an SE is equal to the sum of all possible transitions from an initial state (h,c) to the intermediate state $(h-1,c)$, since there is only one outlet (hot) through which hot cells can leave the SE

$$\lambda_{i,r,\mathrm{hot}} = \sum_{h1=1}^{B}\sum_{c1=0}^{B-h1}\pi_{i,r}(h1,c1) \times \sum_{c3=0}^{c1}\tau_{i,r}(h1,c1,h1-1,c3). \tag{41}$$

The aggregate throughput of the cold outlets is obtained by

$$\lambda_{i,r,\mathrm{cold}} = \sum_{c1=1}^{B}\sum_{h1=0}^{B-c1}\pi_{i,r}(h1,c1) \times \sum_{h3=0}^{h1}\sum_{c3=\max(0,c1-d)}^{c1-1}(c1-c3)\tau_{i,r}(h1,c1,h3,c3). \tag{42}$$

Note that in Eq. (42) cold outlets are indistinguishable. In the equation, $(c1-c3)$ is the number of cells that leave the SE from its cold outlets. The limits of the summations ensure that only the transitions that contribute to the throughput of the cold outlets are considered.

Summing the hot and cold throughputs of an SE gives the overall throughput of that SE as follows:

$$\lambda_{i,r} = \lambda_{i,r,\mathrm{hot}} + \lambda_{i,r,\mathrm{cold}}. \tag{43}$$

Finally, the throughput of stage i is given by

$$\Lambda_i = d^{k-i}\left[\lambda_{i,1} + (d-1)\sum_{r=2}^{i}\lambda_{i,r}d^{r-2}\right]. \tag{44}$$

Eq. (44) applies to all stages including the first stage where $\sum$ becomes irrelevant.

Packet loss. Since there is no cell loss inside the switch, the cell loss probability at any switch input is obtained from the number of cells offered to the first stage and the throughput of that stage

$$\eta = \frac{\rho - \Lambda_1/N}{\rho}, \tag{45}$$

where Λ_1/N is the throughput per link at the first stage.

Delay. The delay of hot and cold outlets of an SE may be calculated using Little's formula by dividing the average queue length by the departure rate of the queue. In a shared buffer SE, each outlet has a logical queue whose length is equal to the number of cells passing through that outlet.

The delay of the hot outlet is calculated as follows:

$$w_{i,r,\mathrm{hot}} = \frac{1}{\lambda_{i,r,\mathrm{hot}}}\sum_{h=0}^{B}\sum_{c=0}^{B-h}h\pi_{i,r}(h,c), \tag{46}$$

where $\lambda_{i,r,\mathrm{hot}}$ is the throughput of the hot outlet, and the summation comprises the average queue length. $w_{i,r,\mathrm{cold}}$ is defined as the average delay of a cell in any one of the logical queues of the cold outlets:

$$w_{i,r,\mathrm{cold}} = \frac{1}{\lambda_{i,r,\mathrm{cold}}}\left(\sum_{h=0}^{B}\sum_{c=0}^{B-h}c\pi_{i,r}(h,c)\right)\frac{1}{d-1} = \frac{1}{(d-1)\lambda_{i,r,\mathrm{cold}}}\sum_{h=0}^{B}\sum_{c=0}^{B-h}c\pi_{i,r}(h,c). \tag{47}$$

Since all $d-1$ cold outlets of $\mathrm{SE}_{i,r}$ have equal throughput, the average logical queue length for any cold outlet of the SE is obtained in Eq. (47) as the average number of cells in the logical queues for all $d-1$ outlets divided by $d-1$. The average delay in $\mathrm{SE}_{i,r}$ is obtained by dividing the sum of the delays at all outlets by d:

$$w_{i,r,av} = \frac{w_{i,r,\mathrm{hot}} + (d-1)w_{i,r,\mathrm{cold}}}{d}. \tag{48}$$

The average delay at stage i is obtained by adding the delays in all SEs at stage i, and then dividing the sum by the number of SEs in the stage (N/d):

$$w_i = \frac{N}{d^i}\left[w_{i,1,\mathrm{av}} + (d-1)\sum_{r=2}^{i}w_{i,r,\mathrm{av}}d^{r-1}\right]\frac{1}{N/d} = \frac{1}{d^{i-1}}\left[w_{i,1,\mathrm{av}} + (d-1)\sum_{r=2}^{i}w_{i,r,\mathrm{av}}d^{r-1}\right]. \tag{49}$$

Finally, the average delay of a cell traversing the Delta switch is obtained by summing the delays in the different stages of the switch:

$$W = \sum_{i=1}^{k} w_i, \tag{50}$$

where k is the number of stages in the network.

4. Simulation study

We have validated the model presented in Section 2 with a simulation study. The same assumptions as made for the analysis apply to the simulation of the network. The following operations are implemented in the simulator:

- At each cycle, a cell is generated with probability ρ (offered load to the switch input). The generated cell is independent of the cells generated in previous cycles and at other input ports. Each cell consists of the following information:
  1. a source tag which identifies the input link at which the packet arrived,
  2. a destination tag denoting the output link to which the cell is destined, and
  3. the current cycle number, used for measurement of the cell delay in the network.
- Simulation results from the first several hundred cycles of the switch operation are ignored to allow the switch to reach a steady-state condition. The simulation program is then allowed to run until the change in the average throughput between consecutive cycles becomes less than 10^{-6}.
- Conflict in the buffers for accessing a particular outlet as well as contention to seize a buffer space in the next stage is resolved using a random number generator with a different seed value from that of the cell generator.

The simulation operates as follows:

1. The cells at the last stage buffers are sent to the output links of the network, and the instantaneous throughput and delay are measured for every link.
2. For each SE at stages $k-1$ to 1:
   - The SE buffers are examined for cells destined to the different outlets of the SE, copies of all cells destined to different outlets are placed in the corresponding outlet lists (forming logical output queues), and the lists are sent to the corresponding inlets of the next stage.
   - If the number of available buffer spaces in an SE is less than the number of cells in the different lists at its inlets, a number of cells equal to the number of available spaces are chosen at random from the available lists. Packets which are not accepted stay in the buffers at the previous stage until they can be forwarded in the subsequent cycles.
3. At the beginning of a clock cycle, a new set of cells are generated at the inputs of stage 1 with probability ρ and hot spot probability f_h. The cells are placed in the first stage buffers if there is any room. If a cell cannot be placed in the first stage buffers, it is discarded, and the cell loss counter is incremented by 1.

5. Numerical results

The normalized throughput of a Delta interconnection for $N = 256$, $d = 2$, $B = 2$, and hot spot values 0 (uniform traffic) and 0.005 is illustrated in Fig. 4. In this figure, the model is quite accurate when the input load is small.

The average delay W of the same switch is shown in Fig. 5. The results from the model are consistent with the simulation results for buffer sizes 2 and 4. As in Fig. 4, the results from the model are close to those of the simulation for small

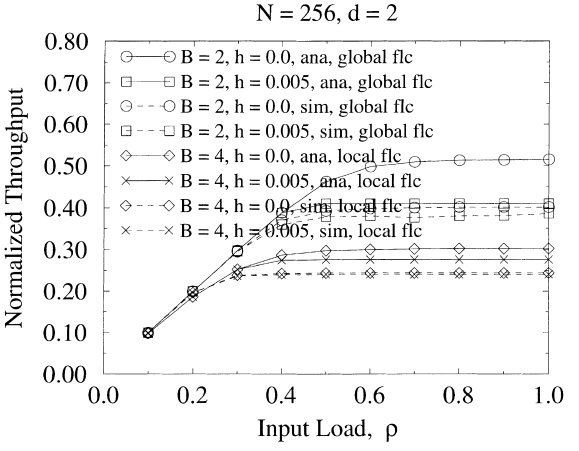

Fig. 4. Normalized throughput versus ρ for $N = 256$, and $d = 2$.

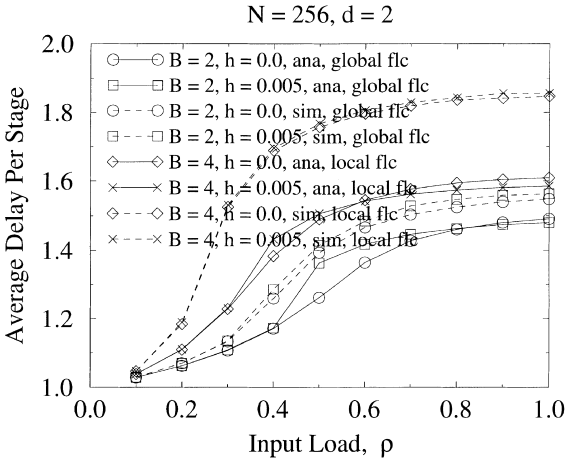

Fig. 5. Average delay versus ρ for $N = 256$, and $d = 2$.

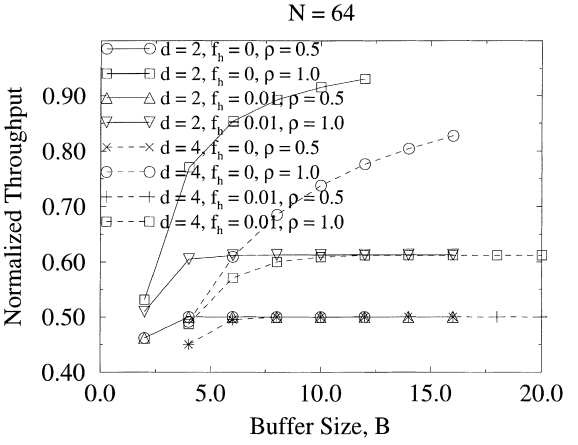

Fig. 6. Normalized throughput versus B for $N = 64$, and global flow control.

input loads. The model predicts lower delay time due to the fact that it ignores the effect of blocked cells which in practice cause higher delays in a network.

Although the throughput of the hot output of a switch increases sharply when the hot spot value increases, the overall throughput of the switch decreases due to the buffer monopolization effect [28] caused by the hot traffic. This situation has a greater impact on the overall performance of a switch when d is large. The reason is that, when a tree saturation [29] takes place in a network, the cold outputs in the type 1 (hot type) SE are most affected by the phenomenon. The effect eases as switch type increases. For a particular switch size N, the smaller the SE size (d) and the greater the number of SE types in the last stage, the less is the effect of the hot spot traffic on the overall throughput of the network. Increasing the buffer size may alleviate the monopolization effect inside the switch under low hot spot values. However, due to the properties of the shared buffer network, increasing the buffer size has very little effect on improving the performance of the switch when hot spot traffic increases. For example, as shown in Fig. 6, buffer sizes greater than 4 have no effect on the throughput of the network for $d = 2$, $f_h = 0.01$, and $\rho = 1.0$. The impact of increasing the buffer size on delay and cell loss under uniform and hot spot traffic is shown in Figs. 7 and 8.

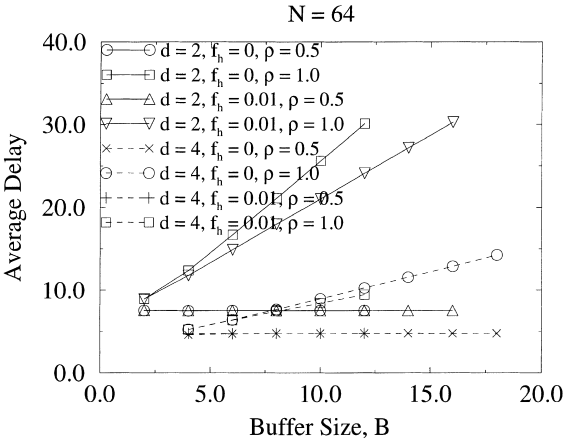

Fig. 7. Average delay versus B for $N = 64$, and global flow control.

A comparison between the model and simulation has been made in Fig. 9 for $N = 256$ and $d = 4$. The results obtained by the model are close to the simulation results under both uniform and hot spot traffic. The small inaccuracy of the model under uniform traffic is due to the fact that the model does not take the time correlation of the blocked packets [30] into account.

Figs. 10 and 11 illustrate hot, cold and total buffer occupancy, expressed as fractions of the buffer spaces occupied by the hot, cold and overall traffic in the first stage SE, respectively. Under

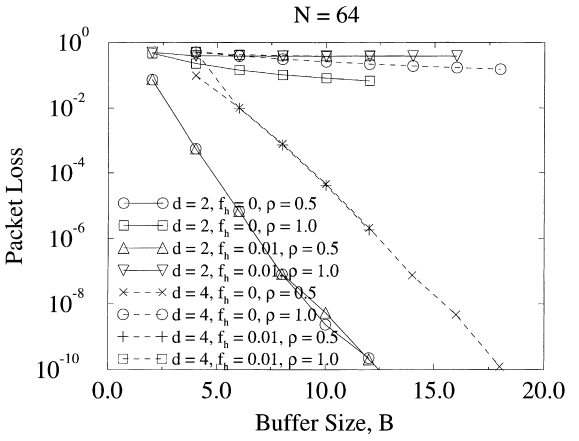

Fig. 8. Packet loss versus B for $N = 64$, and global flow control.

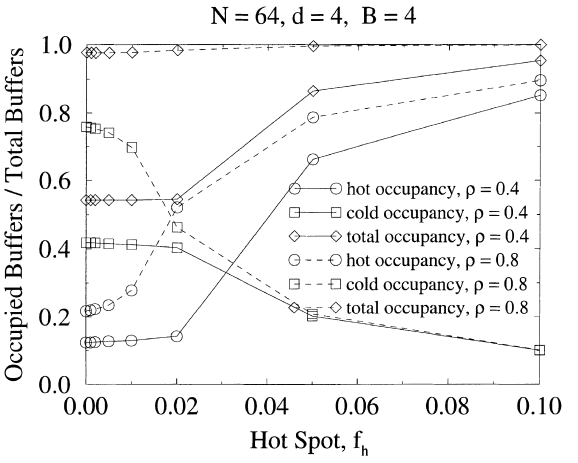

Fig. 10. Ratio of hot, cold, and total buffer occupancy of the first stage SE for $N = 64, d = 4, B = 4$, and global flow control.

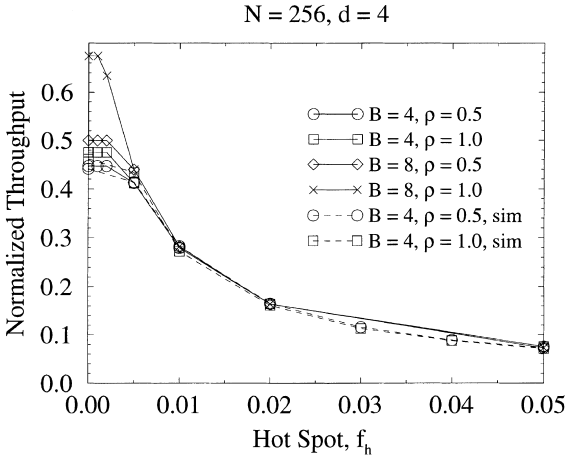

Fig. 9. Normalized throughput versus f_h for $N = 256$, and global flow control.

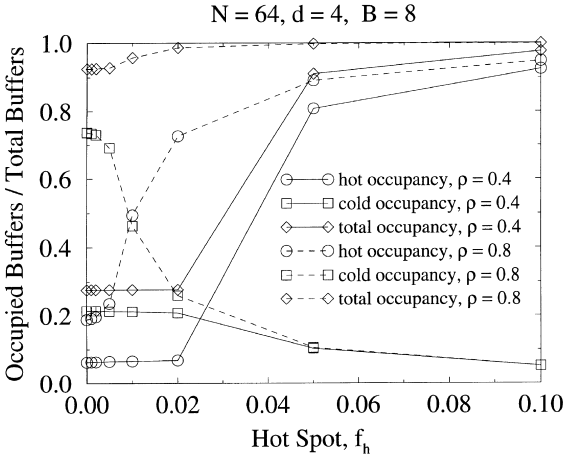

Fig. 11. Ratio of hot, cold, and total buffer occupancy of the first stage SE for $N = 64, d = 4, B = 8$, and global flow control.

uniform traffic, buffer occupancy is proportional to the offered traffic of the network, all outlets taking a fair amount of the total buffer space. However, this proportion changes in favor of the hot traffic, when some hot spot value is introduced. When the hot spot is more than 0.1, the hot traffic saturates the buffers, even under an input load as low as 0.4. Again, increasing the buffer size has little impact on this effect. We have reported similar results for different buffer size and hot spot values in [28]. Better buffer utilization is an advantage of a shared buffer switch which improves the throughput as compared to switches using other buffer disciplines. However, under hot spot traffic, all of the buffers may be exhausted by the hot traffic. Fig. 12 contrasts a shared buffer and an output buffer switch for $N = 64$, and $d = 2$. For a reasonable comparison, we have assumed the same number of buffer spaces (B) per SE for both architectures. The results for the output buffer switch are obtained from a simulation program using a similar methodology to that described in Chapter

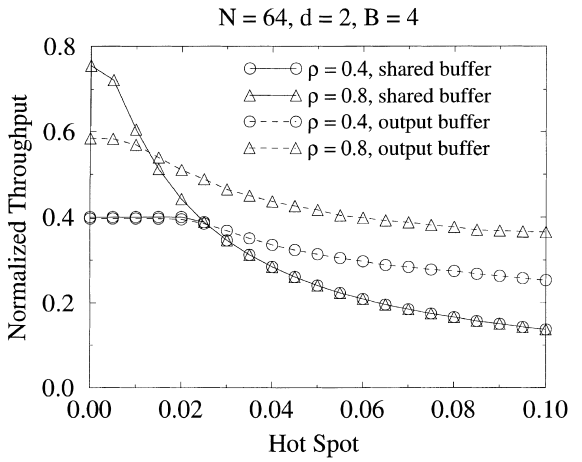

Fig. 12. Comparison of the throughput of shared buffer and output buffer switches for $N = 64$, and $d = 2$.

4. As shown in Fig. 12, a shared buffer switch performs better under uniform traffic or when the hot spot value is small. Under high hot spot values, an output buffer switch performs better. This can be explained as follows. In output buffering, the hot traffic degrades the throughput of the outputs which share the same buffers as a cell destined to them traverses through different stages. However, there are still some outputs that do not share any buffer with the hot output, and therefore are not affected by the hot traffic. Unlike in output buffering, the hot spot traffic in a shared buffer affects all of the outputs, since all of them share the buffers at least at the first stage which, if congested, will degrade the throughput of all non-hot outputs as well.

Buffer monopolization in shared buffer switches can be minimized if a proper buffer management is utilized to limit the maximum number of buffers used by any outlet to some specified value.

6. Conclusion

We have developed an analytical model to study the performance of multistage switches constructed from shared buffer switching elements with an arbitrary SE size and buffer size using either global or local flow control policy. From the model, the throughput, cell delay, and cell loss probability in such switches have been derived, and various numerical results have been illustrated. We have also compared the results obtained from the model and the computer simulation, and they have been found to be in close agreement. The model does not account for the correlation of cells in successive cycles. Hence, a blocked cell in the buffers is treated the same as a new coming packet. This results in predicting higher throughput than simulation. In reality, a blocked cell in an SE always hunts for the same outlet of the SE during successive cycles.

Under uniform traffic or under low hot spot values, a shared buffered switch has better performance in terms of throughput, delay and cell loss as compared to a switch with output buffering. The model can be used by switch designers to study the effect of the different switch parameters on the performance, and optimize the cost/performance ratio of switches with single hot spot distribution.

References

[1] F.M. Chiussi, Y. Xia, V.P. Kumar, Performance of shared memory switches under multicase bursty traffic, IEEE Journal on Selected Areas in Communications 15 (3) (1997) 473–487.

[2] S.F. Oktug, M.U. Caglayan, Design and peformance evaluation of a banyan networks based interconnection structure for ATM switches, IEEE Journal on Selected Areas in Communications 15 (5) (1997) 807–816.

[3] S. Singh, S. Fong, M. Atiquzzaman, An analytical model and performance analysis of shared buffer ATM switches under non-uniform traffic, Computer Systems Science and Engineering 12 (2) (1997) 125–137.

[4] C.Y. Roger Chen, A.S. Almazyad, Performance evaluation of buffered multistage interconnection networks with look-ahead contention resolution scheme, in: Proceedings of the IEEE International Conference on Communications, Seattle, USA, June 1995, pp. 1137–1141.

[5] M. Atiquzzaman, Buffer dimensioning for congestion control in ATM switches, International Journal of Parallel and Distributed Systems and Networks 2 (4) (1999) 217–224.

[6] M. Atiquzzaman, C.K. Chen, Realistic modeling of blocked packets for accurate performance evaluation of multistage ATM switches, IEE Proceedings – Communications 146 (4) (1999) 213–221.

[7] A.K. Choudhury, E.L. Hahne, A new buffer management scheme for hierarchical shared memory switches, Journal of Selected Areas in Communications 5 (5) (1997) 728–735.
[8] H. Yoon, K. Lee, M. Liu, K. Lee, Y. Kim, The knockout switch under nonuniform traffic, IEEE Transactions on Communications 43 (6) (1995) 2149–2156.
[9] L. Bosack, C. Hedrick, Problems in large LANs, IEEE Network 2 (1) (1988) 49–56.
[10] I.I. Makhamreh, Throughput analysis of input-buffered atm switch, IEE Proceedings – Communications 145 (21) (1998) 15–18.
[11] L.T. Wu, Mixing traffic in a buffered Banyan network, in: Proceedings of the Ninth Data Communication Symposium, Whistler Mountain, BC, Canada, September 1985.
[12] H.S. Kim, A. Leon-Garcia, Performance of buffered Banyan networks under nonuniform traffic patterns, IEEE Transactions on Communications 38 (5) (1990) 648–658.
[13] G. Bianchi, A. Pattavina, Architecture and perfrmance of non-blocking atm switches o with shared internal queueing, Computer Networks and ISDN Systems 28 (6) (1996) 835–853.
[14] B. Zhou, M. Atiquzzaman, Efficient analysis of multistage interconnection networks using finite output-buffered switching elements, Computer Networks and ISDN Systems 28 (13) (1996) 1809–1829.
[15] B. Zhou, M. Atiquzzaman, Analysis of window by-pass mechanism in split shared buffer ATM switches, International Journal of Electrical & Electronics Engineers, Australia, 18 (3) (1998) 217–230.
[16] B. Zhou, M. Atiquzzaman, Performance of ATM switch fabric using cross point buffers, Computer Communications 20 (13) (1997) 1146–1159.
[17] T. Lin, L. Kleinrock, Performance analysis of finite-buffered multistage interconnection networks with a general traffic pattern, Performance Evaluation Review 19 (1) (1991) 68–78.
[18] T. Lang, L. Kurisaki, Nonuniform traffic spots (NUTS) in multistage interconnection networks, Technical Report CSD880001, UCLA Computer Science Department, January 1998.
[19] S. Gianatti, A. Pattavina, Performance analysis of shared-buffered Banyan networks under arbitrary traffic patterns, in: IEEE INFOCOM'93, 1993, pp. 943–952.
[20] M.S. Esfahani, M. Atiquzzaman, Performance analysis of shared buffer switches under non-uniform traffic pattern, in: Proceedings of the Australian Telecommunications and Networking Applications Conference, ATNAC'94, Melbourne, Australia, 1994, pp. 283–287.
[21] M. Saleh, M. Atiquzzaman, Accurate modeling of the queueing behavior of shared buffer ATM switches, International Journal of Communication Systems 12 (4) (1999) 287–308.
[22] M. Saleh, M. Atiquzzaman, An accurate performance model of shared buffer ATM switches under hot spot traffic, Computer Communications 22 (6) (1999) 516–522.
[23] M.S. Esfahani, M. Atiquzzaman, Queueing analysis of shared buffer switches for ATM networks, in: Proceedings of the IEEE GLOBECOM'94, San Francisco, USA, December 1994, pp. 1070–1074.
[24] M. Atiquzzaman, M.S. Akhtar, Effect of non-uniform traffic on the performance of unbuffered multistage interconnection networks, in: IEE Proceedings – Computer Digital Techniques, vol. 141, No. 3, May 1994, pp. 169–176.
[25] Y. Jenq, Performance analysis of a packet switch based on single-buffered Banyan network, IEEE Journal on Selected Areas in Communications SAC-16 (6) (1983) 1014–1021.
[26] G. Bianchi, J. Turner, Improved queueing analysis of shared buffer switching networks, in: INFOCOM'93, 1993, pp. 1392–1399.
[27] M.S. Esfahani, M. Atiquzzaman, Analysis of shared multistage networks with hot spot, in: Proceedings of IEEE International Conference on Algorithms and Architectures for Parallel Processing: ICA3APP'95, Brisbane, Australia, April 1995, pp. 799–808.
[28] M.S. Esfahani, M. Atiquzzaman, Buffer occupancy in ATM switches with single hot spot, Electronics Letters 31 (1) (1995) 13–15.
[29] G.F. Pfister, V. Alan Norton, Hot spot contention and combining in multistage interconnection networks, IEEE Transactions on Computers C 34 (10) (1985) 943–948.
[30] B. Zhou, M. Atiquzzaman, Impact of switch architectures on the performance of multistage interconnection networks, in: Proceedings of IEEE TENCON 94, Singapore, 22–26 August 1994, pp. 365–369.

Mahmoud Saleh received his Ph.D. from the Department of Computer Science and Computer Engineering of La Trobe University, Australia. His research interests include B-ISDN and ATM networks.

Mohammed Atiquzzaman received the M.Sc. and Ph.D. degrees in electrical engineering and electronics from the University of Manchester Institute of Science and Technology, UK. Currently he is faculty member in the Department of Electrical & Computer Engineering at University of Dayton, OH. He serves on the editorial boards of *IEEE Communications Magazine*, *Computer Communications* journal, *Journal of Telecommunication Systems* and *Journal of Real Time Imaging*. He has guest edited many special issues of various journals including special issues on *Switching and Traffic Management* and *Optical Networks*, *Systems and Devices* and *IP Telephony* of the IEEE Communications Magazine, *ATM Switching* and *ATM Networks* of the International Journal of Computer Systems Science & Engineering, *Next Generation Internet* in the European Transactions on Telecommunications and *Architectures*, *Protocols And Quality Of Service For The Internet Of The Future* of the European Transactions on Telecommunications. He has also served in the technical program committee of many national and international conferences including IEEE INFOCOM, IEEE

Globecom and IEEE Annual Conference on Local Computer Networks. His current research interests are in Broadband ISDN and ATM networks, multimedia over high speed networks, and switching. He has over 100 refereed publications, most of which can be accessed at http://www.engr.udayton.edu/faculty/matiquzz/. He can be contacted at atiq@ieee.org.